EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
Vol. 4, No. 4, 2011, 467-485
ISSN 1307-5543 – www.ejpam.com

Predictive Subset Selection using Regression Trees and RBF Neural Networks Hybridized with the Genetic Algorithm

Oguz Akbilgic[1,*], Hamparsum Bozdogan[2]

[1] *Department of Quantitative Methods, Istanbul University School of Business Administration, Istanbul, Turkey*
[2] *Department of Statistics, Operations, and Management Science,and Center for Intelligent Systems and Machine Learning (CISML), The University of Tennessee, Knoxville, 37996, USA*

Abstract. In this paper we develop a novel nonparametric predictive subset regression modeling procedure that involves a combination of regression trees with radial basis function (RBF) neural networks hybridized with the genetic algorithm (GA) to carry out the subset selection of the best predictors. We use the information-theoretic measure of complexity (ICOMP) criterion of [5, 6, 7, 8] as our fitness function to choose the best approximating radial basis functions and to choose the best subset of predictors with the GA. To avoid the potential singularities in the design matrix, we combine our model with analytical global ridge regression for regularization. On the other hand, estimation and prediction performance of model also taken into account for best subset chosen.

2000 Mathematics Subject Classifications: 62G08; 62J02; 17-08; 62B10; 62-07

Key Words and Phrases: Model Selection, Subset Selection, Information Criteria, Radial Basis Functions, Neural Networks

1. Introduction

High dimensionality of the independent or the predictor variables in regression models increases the model complexity and that makes the analysis difficult. Data mining techniques help practitioners to overcome such problems. In this frame work, model selection is an important tool to reduce the dimensionality and to measure the model complexity is an an important enterprize to find a subset of predictor variables which represent the underlying relationship between the input or predictor and output or response variables. Although in the literature there are many different model selection criteria that have been proposed and used, most of these criteria are based on Akaike's Information Criterion (AIC), or they are based on some variations of AIC. In contrast to AIC, information complexity ($ICOMP$) type criteria constitute a new class or a new generation model selection criteria.

*Corresponding author.

Email addresses: oguzakbilgic@gmail.com (O. Akbilgic), bozdogan@utk.edu (H. Bozdogan)

In model selection procedures, the chosen model is important as much as the chosen model selection criterion. The assumption of linear relationship between input and output variables can lead us to choose wrong subset of variables. Radial Basis Function Neural Networks (RBF-NN), or what statisticians call nonparametric regression models, seem to be more appropriate in general because RBF-NN does not assume any functional relation between input and output variables. Therefore, combining RBF-NN with some statistical techniques can provide us better results and improve the prediction accuracy in regression modeling.

The idea of combining RBF-NN and Regression Trees (RT) goes back to [12] where how to combine RBF-NN and decision trees are explained. Later, [19] extended this idea to combine RBF-NN with regression and classification trees. Based on the results of [12] and [19], in this paper for the first time, we combine RBF-NN and RT model and we hybridize it with ridge regression, to overcome possible singularity problem on design matrix. We introduce the genetic algorithm (GA) to choose best subset of input variables by scoring the information complexity ($ICOMP$) criterion.

The paper is organized as follows. In Section 2, we present linear models and radial basis function neural networks (RBF-NN) with least squares estimation. Section 3, presents combination of regression trees and RBF-NN. In this section we discuss how to transform the tree nodes into RBFs. In Section 4, we present subset selection of RBFs and state the current problems of forward, backward, combination of forward and backward, and all possible subset selection procedures currently used in the literature. To avoid over-fitting and the potential singularities in the regression design or model matrix, in Section 5, we discuss two main ways of regularization and present global and local ridge regression. We provide several ways of choosing optimal ridge parameters. Section 6 presents several information-theoretic model selection criteria. For space considerations, we restrict the detailed proofs and derivations of these criteria where appropriate. For more details on information criteria, we will refer the readers to [5, 6, 7, 8]. We further provide the derived forms of the model selection criteria in RBF-NN. In Section 7, we present the general background of the genetic algorithm (GA) and its implementation within the RBF-NN. In Section 8, we provide a large scale simulation study using a highly nonlinear simulation protocol where we first choose the best RBF. Then, we carry out a GA subset selection of best predictors and give the regression tree. Following this, we construct the best predictive RBF-NN model based on the best predictors chosen. As an end result, we build the final best fitting RBF-NN model in its open analytical form using the recovered RBF centers, c, radius r, and the regression weights, w. Although the structure of hybrid RBF-NN model represents a very complicated equation, nevertheless, it provides useful information of the structure of the predictive model which is nonlinear. Section 9 concludes the paper.

2. Linear Models and Radial Basis Function Neural Networks (RBF-NN)

2.1. Linear Models

We shall consider supervised learning, or what statisticians call, nonparametric regression problem for a given multi-dimensional data set with the dependent variable y and indepen-

dent (or predictor) variables $x_1, x_2, \ldots, x_m$. We define the general linear model as

$$y = f(w,x) = \sum_{j=1}^{m} w_j h_j(x) = w_1 h_1 + w_2 h_2 + \ldots + w_m h_m, \tag{1}$$

where the regressors, $\left\{h_j(x)\right\}_{j=1}^{m}$, are fixed basis functions (or the transfer functions of the hidden units) of the predictors, $x \in \Re^n$, and $\left\{w_j\right\}_{j=1}^{m}$ are the unknown adaptable coefficients, or weights.

To perform linear regression with this model, we solve the following system of equations:

$$y = Hw + \varepsilon, \tag{2}$$

where y is a vector of $(n \times 1)$ observations on a dependent variable, and H is a $(n \times m)$ design matrix and are responses of m regressors given by

$$H_{(n\times m)} = \begin{bmatrix} h_1(x_1) & h_2(x_1) & \cdots & h_m(x_1) \\ h_1(x_2) & h_2(x_2) & \cdots & h_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(x_n) & h_2(x_n) & h_1(x_1) & h_m(x_n) \end{bmatrix}. \tag{3}$$

In (2), w is a $(n \times 1)$ coefficient vector, and ε is a $(n \times 1)$ vector of random noise term, such that $\varepsilon \sim N(0, \sigma^2 I)$ or equivalently $\varepsilon_i \sim N(0, \sigma^2 I)$, $for\ i = 1, 2, \ldots, n$.

2.2. Radial Basis Functions

The flexibility of f in (1) stems from the fact that we can consider and fit many different radial basis functions (RBFs). RBFs are one possible choice for the hidden unit activation functions in a linear network. The most distinguishing feature of these functions is that they are local, or at least their response decreases monotonically away from a center point. The RBFs are used in function approximation, regularization, noisy interpolation, density estimation optimal classification and clustering, etc. The RBFs, we shall consider, are given below.
Gaussian Kernel (GK):

$$h_j(x) = exp(-\sum_{k=1}^{p} \frac{(x_k - c_{jk})^2}{r_{jk}^2}) \tag{4}$$

Cauchy Kernel (CK):

$$h_j(x) = \frac{1}{1 + exp(-\sum_{k=1}^{p} \frac{(x_k - c_{jk})^2}{r_{jk}^2})} \tag{5}$$

Multiquadric Kernel (MLQK):

$$h_j(x) = \sqrt{1 + exp(-\Sigma_{k=1}^{p} \frac{(x_k - c_{jk})^2}{r_{jk}^2})} \tag{6}$$

Inverse Multiquadric Kernel (IMLQK):

$$h_j(x) = \frac{1}{\sqrt{1 + exp(-\Sigma_{k=1}^{p} \frac{(x_k - c_{jk})^2}{r_{jk}^2})}} \tag{7}$$

2.3. Radial Basis Function Neural Networks

The RBF-NN introduces a mapping or transformation of the n-dimensional inputs nonlinearly to an m-dimensional space and then estimate a model using linear regression. The nonlinear transformation is achieved using m basis functions, each characterized by their center c_j in the (original) input space and a width or radius vector r_j , $j \in \{1, 2, \ldots, m\}$ [19]. In principle, RBFs can be used in any sort of modeling, whether they are linear or nonlinear and for single-layer or multi-layer networks. [20] has shown that RBF-NN possess the property of best approximation.

2.4. Least Squares Estimation

Given a network (or model) in (1) consisting of m RBFs with centers $\left\{c_j\right\}_{j=1}^{m}$ and radii $\left\{r_j\right\}_{j=1}^{m}$ and a training set with p patterns, $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{p}$, the optimal network weights can be found by minimizing the sum of squared errors:

$$SSE = \sum_{i=1}^{p} \left(f(x_i) - y_i\right)^2 \tag{8}$$

and is given by

$$\hat{w} = \left(H^{'} H\right)^{-1} H^{'} y \tag{9}$$

the so called normal equation. Here H is the design matrix, with its elements $H_{ij} = h_j(x_i)$, and $y = \left(y_1, y_2, \ldots, y_p\right)^{'}$ is the p-dimensional vector of training set output values.

3. Combining Regression Trees and RBFNN

3.1. Regression Trees

The basic idea of RT is to partition the input space recursively into two, and approximate the function in each half by the average output value of the samples it contains to refine the subset variable selection [9]. Each split is parallel to one of the axes and can be expressed as an inequality involving of the input components $\left(e.g. x_k > b\right)$. The input space is divided into hyperrectangles organized into a binary tree where each branch is determined by the dimension (k) and boundary (b) which together minimize the residual error between model and data [19]. The root node of the regression tree is the smallest hyperrectangle that will include all of the training data $\left\{x_i\right\}_{i=1}^{p}$. Its size s_k (half–width) and center c_k in each dimension

k are

$$s_k = \frac{1}{2}\left(\underset{i\in S}{max}\left(x_{ik}\right) - \underset{i\in S}{min}\left(x_{ik}\right)\right) \tag{10}$$

$$c_k = \frac{1}{2}\left(\underset{i\in S}{max}\left(x_{ik}\right) + \underset{i\in S}{min}\left(x_{ik}\right)\right) \tag{11}$$

where $k \in K$ is the set of predictor indices, and $S = \{1, 2, \ldots, p\}$ is the set of training set indices. A split of the root node divides the training samples into left and right subsets, S_L and S_R, on either side of a boundary b in one of the dimensions k such that

$$s_L = \{i : x_{ik} \leq b\}, \tag{12}$$

$$s_R = \{i : x_{ik} > b\}. \tag{13}$$

The mean output value on either side of the bifurcation is

$$\overline{y}_L = \frac{1}{p_L}\sum_{i\in S_L} y_i, \tag{14}$$

$$\overline{y}_R = \frac{1}{p_R}\sum_{i\in S_R} y_i, \tag{15}$$

where p_L and p_R are the number of samples in each subset. The mean square error (MSE) is then calculated as in equation (16).

$$MSE(k,b) = \frac{1}{p}\left(\sum_{i\in S_L}\left(y_i - \overline{y}_L\right)^2 + \sum_{i\in S_R}\left(y_i - \overline{y}_R\right)^2\right) \tag{16}$$

The split which minimizes $MSE\,(k,b)$ over all possible choices of k and b is used to create the "children" of the root node and is found by simple discrete search over m dimensions and p observations. The children of the root node are split recursively in the same manner and the process terminates when every remaining split creates children containing fewer than p_{min} samples, which is a parameter of the method. The children are shifted with respect to their parent nodes and their sizes reduced in the $k-th$ dimension.

RT can both estimate a model and indicate which components of the input vector most relevant to the modeled relationship. Dimensions which carry the most information about the output tend to split earliest and most often [19].

3.2. Transforming Tree Nodes Into RBFs

The regression tree contains a root node, some nonterminal nodes (having children) and some terminal nodes (having no children). Each node is associated with a hyperrectangle of input space having a center c and size s as described above. The node corresponding to the largest hyperrectangle is the root node and that is divided up into smaller and smaller pieces progressing down the tree. To transform the hyperrectangle into different basis kernel RBFs

we use its center c as the RBF center and its size s, scaled by a parameter α as the RBF radius given by

$$r = \alpha s. \tag{17}$$

The scalar α has the same value for all nodes (Kubat, 1998), and it is another parameter of the method. One can use $\alpha = \sqrt{2}\alpha_K^{-1}$ where α_K is the Kubat's parameter [12, 19].

4. Subset Selection of RBFs and Current Problems

After the tree nodes are transformed into RBFs, the next step of the method is to carry out a subset selection of variables to be included in the model to choose the best fitting subset(s). Current standard techniques for variable selection include:

- Forward selection: The basis kernel RBFs are added until-over-fitting occurs
- Backward elimination: The basis kernel RBFs are pruned until over-fitting is prevented.
- A combination of the two: Two forward selection steps followed by one backward elimination step.
- All possible subset selection: Full combinatorial search.

There are some problems with these techniques. Both forward and backward procedures can not deal with the collinearity in the predictor variables. Major criticisms on the forward, backward, and stepwise selection are that, little or no theoretical justification exists for the order in which variables enter or exit the algorithm [3, 25]. On the other hand, stepwise searching rarely finds the overall best model or even the best subsets of a particular size [17, 10, 18]. Stepwise selection, at the very best, can only produce an "adequate" model. All possible subset selection is a fail proof method, but it is not computationally feasible. It takes too much time to compute and it is costly. For 20 predictor variables, for the usual subset regression model, total number of possible models we need to evaluate is: $2^{20} = 1,048,576$. The regression trees can automatically determine the relevance of the variables. But they still tend to overfit the model because the regression tree method does not discard any of the predictor variables out of the models. In this case, we have $2^{2^{20}} = 2^{1,048,576}$ possible models to evaluate and to choose from.

Other major problems of these standard techniques are, over-fitting, ill-conditioned design matrix, high collinearity in the predictor variables and, computational complexity, etc. In this case what we need is an intelligent hybrid modeling between:

- Any complex modeling problems such as regression trees with RBF-NN models.
- A clever model choice criteria such as the information complexity;
- Fast and efficient stochastic search algorithms such as the genetic algorithms (GA), and
- Hybridization of GA with combinatorial all possible subset selection.

5. Regularization: Ridge Regression

There are two main ways to avoid over-fitting and to avoid the potential singularities in the design matrix. The first way, regularization [23, 2], reduces the "number of good parameter measurements" [16] in a large full saturated model by adding a weight penalty term to the minimization criterion. We introduce the regularization by using global ridge regressions to avoid the potential singularities in the model matrix.

Second way to avoid over-fitting is to explicitly limit the complexity of the network by allowing only a subset of the variables using information criteria to determine the parsimonious networks and best subset of predictors. In this paper, we use not only ridge regression but also subset selection to avoid over-fitting and singularity problems.

5.1. Global Ridge Regression

In the global ridge regression to counter the effects of over-fitting, a roughness penalty term is added to the sum of squared errors to produce the cost function;

$$C(w,\lambda) = \sum_{i=1}^{p} \left(f(x_i) - y_i\right)^2 + \lambda \sum_{i=1}^{m} w_j^2 = \varepsilon'\varepsilon + w'w \tag{18}$$

which is minimized to find a weight vector which is more robust to noise in the training set. The optimal weight vector for global ridge regression is

$$\hat{w} = \left(H^{'}H + \lambda I_m\right)^{-1} H^{'}y \tag{19}$$

where I_m is the m dimensional identity matrix.

5.2. Local Ridge Regression

We generalize the global ridge regression to attach a separate regularization parameter to each basis function by using the cost function

$$C(w,\lambda) = \sum_{i=1}^{p} \left(f(x_i) - y_i\right)^2 + \sum_{i=1}^{m} \lambda_j w_j^2. \tag{20}$$

Leading to the optimal weight of

$$\hat{w} = \left(H^{'}H + \Lambda\right)^{-1} H^{'}y, \tag{21}$$

where $\Lambda = diag\{\lambda_j\}_{j=1}^{m}$ is a diagonal regularization parameter matrix.

5.3. Choosing the Optimal Ridge Parameter

There is much controversy as to how to choose the ridge parameter λ. Several authors have proposed analytical procedures for choosing the optimal parameter λ. Some of these are:

- Hoerl, Kennard & Baldwin (HKB) [11] approach to choosing λ

$$\hat{\lambda}_{HKB} = \frac{ms^2}{\hat{w}_{LS}^{'}\hat{w}_{LS}} \tag{22}$$

where $m = k$, the number of predictors not including the intercept term, n is the number of observations, s^2 is the estimated error variance using k predictors so that

$$s^2 = \frac{1}{(n-k+1)}\left(y - H\hat{w}_{LS}\right)^{'}\left(y - H\hat{w}_{LS}\right) \tag{23}$$

and $\hat{w}_{LS}$ is the estimated coefficient vector obtained from a no-constant model given by

$$\hat{w}_{LS} = \left(H'H\right)^{-1}H'y. \tag{24}$$

- Lawless and Wang [14] suggested that

$$\hat{w}_{LS} = \frac{ms^2}{\sum_{j=1}^{k}\hat{w}_j^2\lambda_j} \tag{25}$$

as an estimator of $\hat{\sigma}^2/\hat{\sigma}_w^2$ based on Bayesian argument.

- Empirical Bayes method of determining λ proposed by Sclove [22]

$$\hat{\lambda}_s = \frac{\hat{\sigma}^2}{\hat{\sigma}_w^2} \tag{26}$$

where

$$\hat{\sigma}^2 = \frac{1}{n}y'\left[I - H\left(H'H\right)^{-1}H'\right]y \tag{27}$$

is the estimated residual variance and

$$\hat{\sigma}_w^2 = \frac{y'y - n\hat{\sigma}^2}{tr\left(H'H\right)}. \tag{28}$$

6. Information Theoretic Model Selection Criteria

For the model seletion, we use information theoretic measure of complexity (ICOMP) criteria of [5, 6, 7, 8] function to choose the best fitting basis kernel RBFs, and the best subset of predictors with the hybridized GA with regularization of the regression trees and RBF networks.

The complexity of a nonparametric regression model increases with the number of independent and adjustable parameters, also termed effective degrees of freedom, in the model. According to the qualitative principle of Occam's Razor, we need to find the simplest model that fits the observed data. We need to provide a trade off between how well the model fits the data and the model complexity.

The derived forms of information criteria used to evaluate and compare different horizontal and vertical subset selection in the genetic algorithm (GA) for the regularized regression trees and RBF networks model given by (2) under the assumption: $\varepsilon \sim N\left(0, \sigma^2 I\right)$ or equivalently $\varepsilon_i \sim N\left(0, \sigma^2\right)$ or $i = 1, 2, \ldots, n.$ are defined as follows.

1. Several forms of ICOMP Based on Information Complexity Measures [5, 6, 7, 8]: One of the general forms of ICOMP is an approximation to the sum of two Kullback-Leibler (KL) [13] distances.

   - For general multivariate normal linear or nonlinear structural models, suppose $C_1\left(\hat{\Sigma}_{model}\right)$ is approximated by the complexity of the inverse-Fisher information matrix (IFIM) $C_1\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\right)$, then we define ICOMP(IFIM) as

   $$ICOMP(IFIM) = -2logL\left(\hat{\theta}\right) + 2C_1\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\right) \tag{29}$$

   $C_1(\cdot)$ is a maximal information theoretic measure of complexity of IFIM of a multivariate normal distribution given by

   $$C_1\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\right) = \frac{s}{2}logL\left(\frac{tr\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\right)}{s}\right) - \frac{1}{2}log\mid\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\mid \tag{30}$$

   where $s = dim\left(\hat{\mathscr{F}}^{-1}\right) = rank\left(\hat{\mathscr{F}}^{-1}\right)$. For the regression trees and RBF networks, the estimated inverse Fisher information matrix (IFIM) is given by

   $$\widehat{Cov}\left(\hat{w}, \hat{\sigma}^2\right) = \hat{\mathscr{F}}^{-1} = \begin{bmatrix} \hat{\sigma}^2\left(H'H\right)^{-1} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{4} \end{bmatrix}, \tag{31}$$

   where

   $$\hat{\sigma}^2 = \frac{\left(y - H\hat{w}\right)'\left(y - H\hat{w}\right)}{n}. \tag{32}$$

   Then, ICOMP(IFIM) using the definition, becomes:

   $$ICOMP(IFIM) = nln(2\pi) + nlogL\left(\hat{\sigma}^2\right) + n + 2C_1\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\right) \tag{33}$$

   where the entropic complexity

   $$\begin{aligned} C_1\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}_m\right)\right) &= (m+1)log\left[\frac{tr\hat{\sigma}^2\left(H'H\right)^{-1} + \frac{2\hat{\theta}^4}{4}}{m+1}\right] \\ &- \frac{1}{2}log\mid\hat{\sigma}^2\left(H'H\right)^{-1}\mid + log\left(\frac{2\hat{\sigma}^4}{4}\right) \end{aligned} \tag{34}$$

   We can also define ICOMP for misspecified models.

- ICOMP under Misspecification:

$$\begin{aligned} ICOMP(IFIM)_{Misspec} &= -2lnL\left(\hat{\theta}\right) + 2C_1\left(\widehat{Cov}\left(\hat{\theta}\right)_{Misspec}\right) \\ &= nln(2\pi) + nln\left(\hat{\sigma}^2\right) + n + 2C_1\left(\widehat{Cov}\left(\hat{\theta}\right)_{Misspec}\right) \end{aligned} \tag{35}$$

where

$$\widehat{Cov}\left(\hat{\theta}\right)_{Misspec} = \hat{\mathscr{F}}^{-1}\hat{R}\hat{\mathscr{F}}^{-1} \tag{36}$$

is a consistent estimator of the covariance matrix $Cov\left(\theta_k^*\right)$ for

$$\hat{\mathscr{F}}^{-1} = \begin{bmatrix} \hat{\sigma}^2\left(H'H\right)^{-1} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{4} \end{bmatrix}, \text{ and } \hat{R} = \begin{bmatrix} \frac{1}{\hat{\sigma}^4}H'D^2H & H'1\frac{s_k}{2\hat{\sigma}^3} \\ \left(H'1\frac{s_k}{2\hat{\sigma}^3}\right)' & \frac{(n-m)(Kt-1)}{4\hat{\sigma}^4} \end{bmatrix}.$$

This is often called the "sandwich covariance" or "robust covariance" estimator, since it is a correct variance regardless whether of the assumed model is correct or not. When the model is correct we get $\hat{\mathscr{F}} = \hat{R}$, and the formula reduces to the usual inverse Fisher information matrix $\hat{\mathscr{F}}^{-1}$ [24]. Note that this covariance matrix takes into account presence of skewness and kurtosis which is not possible with AIC, and MDL/SBC.

2. Akaike's Information Criterion (AIC) [1]:

$$AIC(m) = nln(2\pi) + nln\left(\frac{\left(y - H\widehat{w}\right)'\left(y - H\widehat{w}\right)}{n}\right) + n + 2(m+1) \tag{37}$$

3. Schwartz Bayesian (SBC) criterion [21]:

$$SBC(m) = nln(2\pi) + nln\left(\frac{\left(y - H\widehat{w}\right)'\left(y - H\widehat{w}\right)}{n}\right) + n + mlog(n) \tag{38}$$

4. Consistent Akaike's Information Criterion using Fisher Information (CAICF) [4]:

$$\begin{aligned} CAICF(m) &= nln(2\pi) + nln\left(\frac{\left(y - H\widehat{w}\right)'\left(y - H\widehat{w}\right)}{n}\right) + n \\ &+ 2(m+1) + log\left|\mathscr{F}\left(\hat{\theta}_k\right)\right| \end{aligned} \tag{39}$$

where $\mathscr{F}\left(\hat{\theta}_k\right)$ is the Fisher information matrix at the parameter estimation $\hat{\theta}_k$.

7. Genetic Algorithm for Subset Selection

The genetic algorithm (GA) is a stochastic or probabilistic search algorithm that employs natural selection and genetic operators. A GA treats information as a series of codes on a binary string, where each string represents a different solution to a given problem. It follows the principles first laid down by Charles Darwin of survival of the fittest. The algorithm searches within a defined search space to solve a problem. It has outstanding performance in finding the optimal solution for problems in many different fields.

Recall that the regularized regression tree and RBF networks model given by 1, the GA is used to find the best or nearly best subset of predictors from the data.

7.1. Implementation of the GA

The GA is implemented using the following steps:

1. Implementing a genetic coding scheme: The first step of the GA is to represent each subset model as a binary string. A binary code of 1 indicates presence and a 0 indicating absence. Every string is of the same length, but contain different combinations of predictor variables. For a data set with $k = 6$ predictors with a constant, following string represents a model including constant, and input variables x_2, x_3, and x_6.

$$\begin{array}{ccccccc} 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ x_0 & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array}$$

2. Generating an initial population of the models: The initial population consists of randomly selected models from all possible models. We have to choose an initial population of size N. Our algorithm allows one to choose any population size. The best population size to choose depends on many different factors and requires further investigation.

3. Using a fitness function to evaluate the performance of the models in the population: A fitness function provides a way of evaluating the performance of the models. We use the $ICOMP$ information criteria defined in the previous section as the fitness function. In general, the analyst has the freedom of using any appropriate model selection criterion as the fitness functions.

4. Selecting the parents models from the current population: This step is to choose models to be used in the next step to generate new population. The selection of parents' models is based on the natural selection. That is, the model with better fitness value has greater chance to be selected as parents. We calculate the difference:

$$\Delta ICOMP_{(i)}(IFIM) = ICOMP(IFIM)_{Max} - ICOMP(IFIM)_i = Range \tag{40}$$

for $i = 1, 2, \ldots, N$, where N is the population size. Next, we average these differences; that is, we compute

$$\overline{\Delta ICOMP(IFIM)} = \frac{1}{N}\sum_{i=1}^{n} \Delta ICOMP_{(i)}(IFIM) \tag{41}$$

Then the ratio of each model's difference value to the mean difference value is calculated. That is, we compute

$$ICOMP_{Ratio} = \frac{\Delta ICOMP_{(i)}(IFIM)}{\overline{\Delta ICOMP(IFIM)}} \tag{42}$$

This ratio is used to determine which models will be included in the mating pool. The chance of a model being mated is proportional to this ratio. In other words, a model with a ratio of two is twice as likely to mate as a model with a ratio of one. The process of selecting mates to produce offspring models continues until the number of offsprings equals the initial population size. This is called the proportional selection or fitting.

5. Produce offspring models by crossover and mutation process: The selected parents are then used to generate offsprings by performing crossover and/or mutation process on them. Both the crossover and mutation probability is determined by the analyst. A higher crossover probability will on one hand introduce more new models into the population in each generation, while on the other hand remove more of the good models from the previous generation. A mutation probability is a random search operator. It helps to jump to another search space within the solutions' scope. [15] states that mutation should be used sparingly because the algorithm will become little more than a random search with a high mutation probability. There are several different ways of performing the crossover. These are single point crossover, two-point crossover, and uniform crossover, etc.

8. Simulation Studies

In this section, we report our computational results on a simulated data set using hybrid RBF-NN approach between the regression trees RBF networks with regularization, the GA and $ICOMP(IFIM)_{Misspec}$. In our numerical example, we use different basis kernels, Gaussian kernel (GK), Cauchy kernel (CK), Multiquadric kernel (MLQK), and Inverse Multiquadric kernel (IMLQK). On the other hand, to choose the optimal ridge parameter λ for the regularization, we use Hoerl, Kennard & Baldwin (HKB) method under four different model selection criteria. Namely, we use AIC, SBC, $CAICF$, and $ICOMP(IFIM)_{Misspec}$. We define regression tree parameters; p_{min} is integer value of 10% of training data sample size, α parameter is 2 or 4 whichever fits better. The GA parameters are: number of generations is 15, population size is 10, crossover type is uniform, probability of crossover is 0.5, probability of mutation is 0.1, and elitist rule is used for optimization.

To carry out a subset selection of variables, we consider the following Monte Carlo simulation protocol. We draw n $U(0,1)$ random numbers and include a model in the mating pool each time when one of the random numbers falls within its bin. Since better models have wider bins, we expect members of the current generation with better model selection criteria scores to be over-represented in the mating pool. This fulfills the natural selection role of the *GA*. The mating pool determined in this way is subjected to a crossover process that deter-

mines the subset regression models included in the next generation. We generate 7 predictor variables by using a constant multiple of uniform random variables between 0 and 1.That is:

$$\begin{array}{|l|}\hline x_1 = 1 \times U(0,1) \\ x_2 = 2 \times U(0,1) \\ x_3 = 3 \times U(0,1) \\ x_4 = 4 \times U(0,1) \\ x_5 = 5 \times U(0,1) \\ x_6 = 6 \times U(0,1) \\ x_7 = 7 \times U(0,1) \\ \hline \end{array}$$

By using some predictors of the model data matrix $X = \left[x_1, x_2, x_3, x_4, x_5, x_6, x_7\right]$, the output or the response variable is generated using following functional relationship:

$$y = 10sin\left(\pi x_1 x_2\right) + 20\left(x_3 - 0.5\right)^2 + 10x_4 + \varepsilon \tag{43}$$

where $\varepsilon \sim N(0,1)$. Note that in this simulation protocol the first three variables are nonlinear, the next is linear to output, then last 3 variables have no effect on the response y. Therefore, true model includes the regressors x_1,x_2, x_3 and, x_4.

8.1. Simulation Study 1

In the first phase of the simulation study, we choose the best kernel function for hybrid RBF model transfer functions according to their model selection performance. To realize this objective, we run 100 simulations using different sample sizes, $n = 50, 100, 250$ and 500, respectively, and then construct Hybrid RBF model with different kernel functions including Gaussian, Cauchy, Multiquadratic, and Inverse Multiquadratic. The true model selection percentages, according to $ICOMP(IFIM)_{Misspec}$ criterion, are summarized in the Table 1. Looking at the results in Table 1, we see that Gaussian kernel function performs the best as

Table 1: Comparison of kernel functions' performances.

Kernel Function	Sample Size			
	50	100	250	500
Gauss	26%	49%	71%	89%
Cauchy	19%	47%	71%	74%
Multiquadratic	13%	25%	68%	87%
Inverse Multiquadratic	17%	45%	70%	78%

compared to other kernel functions for this simulated model.

8.2. Simulation Study 2

The second phase of the simulation study is to compare the performance of the hybrid RBF model approach with that of the classical linear regression model. Our simulation set

up is the same as before. We run the simulation 100 times with different sample sizes: $n = 50, 100, 250$ and 500, respectively and score different model selection criteria AIC, BIC, $CAICF$, $ICOM(IFIM)_{Misspec}$ under the proposed hybrid RBF and the classic linear regression model. Table 2 summarizes the percent hit ratios of the true model.

Table 2: Comparison of Proposed Model and Linear Regression Model.

	Hybrid RBF Model				Linear Regression Model			
n	50	100	250	500	50	100	250	500
AIC	17	58	78	87	10	12	3	0
SBC	24	64	80	90	6	14	17	7
CAICF	19	50	84	87	14	24	45	24
$ICOMP(IFIM)_{Misspe}$	26	49	71	89	22	33	13	1

It is clear from Table 2 that hybrid RBF model is superior to the linear regression model in terms of model selection results. Hybrid RBF model selects the true model with high frequency as the sample size increases. Considering the highly nonlinear relationship between input and output variables, hybrid RBF model performs better in terms of model selection based on all the information criteria. The poor performance of linear regression model on a simulated true Freidman model which has nonlinear structure is not so surprising since the linear regression does not take model misspecification into account and can not handle the singularity problem in the design matrix, $H^{'}H$. On the other hand, due to function approximation and implicit smoothing properties of the hybrid RBF approach guards us from model misspecification as shown in our simulation results in terms of its outstanding performance.

8.3. Simulation Study 3

Third and last phase of our simulation study is to determine the estimation and prediction success of hybrid RBF model using the same simulation protocol as above. We generate training data with sample sizes: $n = 50, 100, 250$ and 500. We use 20 observations of the test data for each. First, we learn model parameters from test data and then we predict our results from the data using parameters determined from the training data. Table 3 gives the training and testing errors in two ways. We also report the root mean square error ($RMSE$) and root mean square percentage error ($RMSPE$). Figures 1 and 2 show that hybrid RBF model fits the data very well not only for training data but also for test data. This aspect can be an evidence to claim that hybrid RBF model learns the relationship within the regression data set considered.

Although the structure of hybrid RBF model represents a very complicated equation, we can build the final best fitting RBF-NN model in its open analytical form using the recovered RBF centers, c, radius r, and the regression weights, w. In (44) we show the constructed hybrid RBF model obtained for sample size $n = 250$ from the the generated regression tree which is shown in Figure 3.

Table 3: Estimation and Prediction Performance of Hybrid RBF Model.

	Error Type			
Sample Size	*RMSE*		*RMSPE*	
Train-Test	Train	Test	Train	Test
$50-20$	9.89	12.39	7.27	3.40
$100-20$	8.08	8.18	11.64	5.09
$250-20$	9.31	10.83	5.33	2.07
$500-20$	8.48	7.90	5.20	2.38

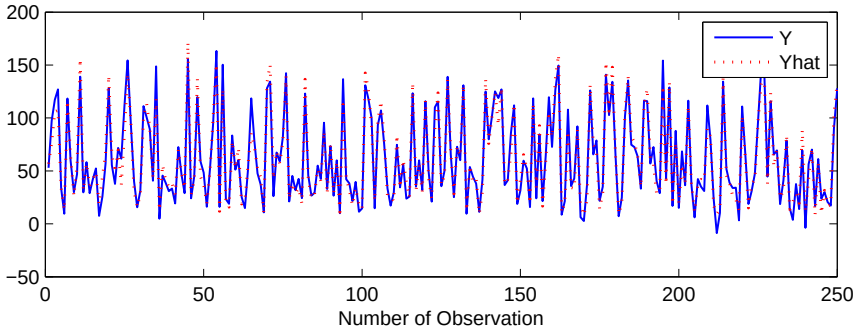

Figure 1: Observed and estimated values of Y for training data.

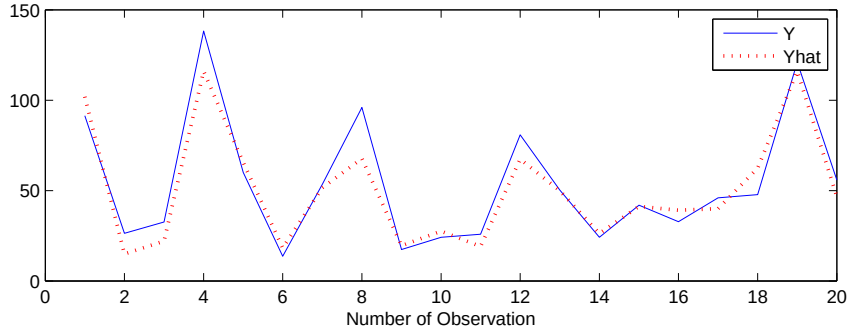

Figure 2: Observed and estimated values of Y for test data.

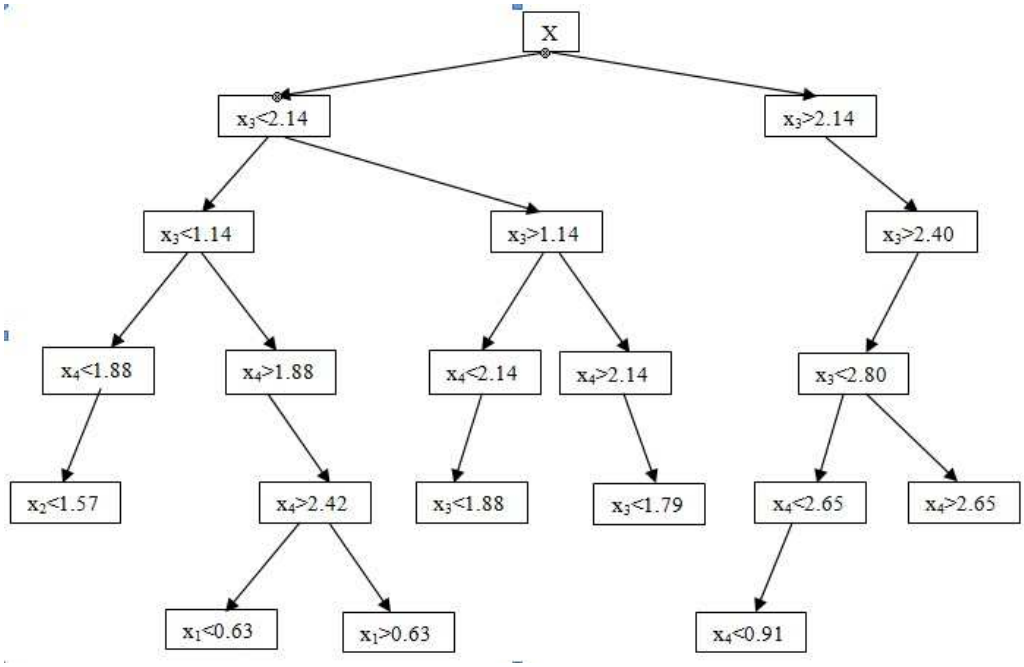

Figure 3: Regression tree developed for $n = 250$.

$$
\begin{aligned}
y &= 10sin\left(\pi x_1 x_2\right) + 20\left(x_3 - 0.5\right)^2 + 10x_4 + \varepsilon \qquad (44)\\
&\approx 483.56exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-1.50}{2.99}\right]^2 + \left[\frac{x_4-2.00}{3.97}\right]^2\right)\\
&- 405.34exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-1.07}{2.14}\right]^2 + \left[\frac{x_4-2.00}{3.97}\right]^2\right)\\
&+ 15.68exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-0.72}{1.43}\right]^2 + \left[\frac{x_4-2.00}{3.97}\right]^2\right)\\
&- 53.74exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-0.72}{1.43}\right]^2 + \left[\frac{x_4-0.95}{1.87}\right]^2\right)\\
&- 50.78exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-1.79}{0.70}\right]^2 + \left[\frac{x_4-1.07}{1.12}\right]^2\right)\\
&+ 13.53exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-0.66}{0.44}\right]^2 + \left[\frac{x_4-1.07}{1.12}\right]^2\right)\\
&+ 69.05exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-0.72}{1.43}\right]^2 + \left[\frac{x_4-2.79}{0.74}\right]^2\right)\\
&- 72.05exp\left(-\left[\frac{x_1-0.31}{0.63}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-0.72}{1.43}\right]^2 + \left[\frac{x_4-2.79}{0.74}\right]^2\right)\\
&- 38.00exp\left(-\left[\frac{x_1-0.81}{0.36}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-0.72}{1.43}\right]^2 + \left[\frac{x_4-2.79}{0.74}\right]^2\right)\\
&- 53.03exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-2.60}{0.39}\right]^2 + \left[\frac{x_4-1.33}{2.63}\right]^2\right)\\
&+ 26.67exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-2.60}{0.39}\right]^2 + \left[\frac{x_4-3.31}{1.33}\right]^2\right)\\
&+ 30.32exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-2.60}{0.39}\right]^2 + \left[\frac{x_4-0.46}{0.90}\right]^2\right)\\
&- 9.19exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-1.00}{1.99}\right]^2 + \left[\frac{x_3-1.97}{0.35}\right]^2 + \left[\frac{x_4-3.06}{1.84}\right]^2\right)\\
&+ 9.03exp\left(-\left[\frac{x_1-0.49}{0.99}\right]^2 + \left[\frac{x_2-0.76}{1.52}\right]^2 + \left[\frac{x_3-1.20}{0.47}\right]^2 + \left[\frac{x_4-0.95}{1.87}\right]^2\right)
\end{aligned}
$$

9. Conclusions

In this paper, we have tackled a very important and common problem in statistical analysis of predictive regression modeling. That is, we showed how to select a best subset of variables in a regression model using the genetic algorithm (GA). In this context, we scored several AIC and $ICOMP$-type criteria to evaluate the hybrid RBF model, RBF-NN combined with regression trees using ridge regression regularization. We used a highly nonlinear simulation protocol that shows the nonlinear functional relationship between input and output variables and that of some redundant variables. Simulation results show that Gaussian kernel function is the best choice for hidden unit transfer function of hybrid RBF-NN. On the other hand, model selection performance of hybrid RBF-NN model is much more superior than the linear regression model. In fact the usual standard linear regression model fails miserably when the data exhibits highly nonlinear structure. The success of hybrid RBF-NN model using model selection criteria as a fitness function consistently improves as the sample size increases. Finally, estimation and prediction performance of hybrid RBF-NN models is measured with respect to $RMSE$ and $RMSPE$ using the best subset of predictor variables chosen. Our results show that, hybrid RBF-NN model is quite adoptive to handle highly nonlinear relationships between the predictor and response variables in regression modeling.

It would be interesting to extend this work to the multivariate case where we have more than one response variable. This work within RBF-NN modeling framework has not been carried out before. We intend to pursue this avenue in a future research initiative.

ACKNOWLEDGEMENTS The first author extends his thanks to The Scientific and Technological Research Council of Turkey (TUBITAK) for their support for Young Researchers Award. The first author is grateful to Prof. Dr. Bozdogan for giving this problem and sharing his initial results on RBF-NN. Without his supervision and guidance this research would not have been possible. Therefore, the first author is gratefully acknowledges Prof. Dr. Bozdogan's guidance.

References

[1] H Akaike. Information theory and an extension of the maximum likelihood principle. In B Petrox and F Csaki, editors, *Second International Symposium on Information Theory.*, pages 267–281, Budapest, 1973. Academiai Kiado.

[2] C Bishop. Improving the generalization properties of radial basis function neural networks. *Neural Computation*, 3:579–588, 1991.

[3] D Boyce, A Fahri, and R Weischedel. *Optimal Subset Selection: Multiple Regression, Independence, and Optimal Network Algorithms Extension.* Springer Verlag, New York, 1974.

[4] H Bozdogan. Model selection and Akaike's information criterion (AIC): The general theory and it's analytical extension. *Journal of Mathematical Psychology*, 52:345–370, September 1987.

[5] H Bozdogan. Icomp: A new model-selection criteria. In H.H Bock, editor, *Classification and Related Methods of Data Analysis.* 1988.

[6] H Bozdogan. Mixture-model cluster analysis using a new informational complexity and model selection criteria. In H Bozdogan, editor, *Multivariate Statistical Modeling, Vol. 2, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, pages 69–113. Kluwer Academic Publishers, The Netherlands, Dordrecht, 1994.

[7] H Bozdogan. Akaike's information criterion and recent developments in informational complexity. *Journal of Mathematical Psychology*, 44:62–91, March 2000.

[8] H Bozdogan. Intelligent statistical data mining with information complexity and genetic algorithms. In H Bozdogan, editor, *Statistical Data Mining and Knowledge Discovery*, pages 15–56. Chapman and Hall/CRC, Boca Raton, Florida, 2004.

[9] L Breiman, J Freidman, J C Stone, and R Olsen. *Classification and Regression Trees*. Chapman and Hall, 1984.

[10] R Hocking. Developments in linear regression methodology: 1959-1982. *Technometrics*, 25:219–230, 1983.

[11] A Horel, R Kennard, and K Baldwin. Ridge regression: Some simulations. *Communications in Statistics*, 4:105–123, 1975.

[12] M Kubat. Decision trees can initialize radial basis function networks. *Transactions on Neural Networks*, 9:813–821, 1998.

[13] A Kullback and R Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[14] J Lawless and P Wang. A simulation study if ridge and other regression estimators. *Communications in Statistics*, A5:307–323, 1975.

[15] C Lin and C Lee. *Neural Fuzzy Systems; A Neuro-Fuzzy Synergism to Intelligent Systems*. Prentice Hall P T R, New Jersey, USA, 1996.

[16] D MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4:448–472, 1992.

[17] N Mantel. Why stepdown procedures in variables selection. *Technometrics*, 12:591–612, 1970.

[18] L Moses. *Think and Explain with Statistics.* Addison-Wesley, MA, 1986.

[19] M Orr. Combining regression trees and rbfs. *International Journal of Neural Systems*, 10:453–465, 2000.

[20] T Poggio and F Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science, New-Series*, 247:978–982, 1990.

[21] G Schwartz. Estimating the dimension of model. *Annals of Statistics*, 6:461–464, 1978.

[22] S Sclove. Least squares with random regression coefficient. Technical report, Department of Economics, Stanford University, 1973.

[23] A Tikhonov and V Arsenin. *Solutions of ill-posed problems.* Wiley, 1977.

[24] H White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1982.

[25] L Wilkinson. *SYSTAT: The System for Statistics.* SYSTAT, Evanston, IL, 1989.